Die Datenbank Gesprochenes Deutsch –

Archivierung, Dokumentation und Erschließung des Deutschen Spracharchivs (DSAv)

von Peter Wagener

Zur Beschreibung und Untersuchung der geschriebenen Realisationsformen einer Sprache kann der Linguist auf größere Mengen geschriebenen Materials zurückgreifen. Ihm stehen Bibliotheken mit gedruckten Medien und zunehmend auch Datenbanken mit elektronisch archivierten und recherchierbaren Daten im Internet oder auf elektronischen Datenträgern zur Verfügung.

Anders der Linguist, der sich mit gesprochener Sprache beschäftigt. Er muss sich seinen flüchtigen, an Ort, Zeit und Situation gebundenen einmalig produzierten Gegenstand erst in einem anderen Medium verfügbar machen, um ihn reproduzieren zu können und für Beschreibung und Analyse zur Verfügung zu haben. Nur behelfsweise bedienen die Sprachforscher sich dafür des Mittels der Mitschrift, um das Gesprochene in Schriftform zu überführen.[1] Es ist sicher kein Zufall, dass sich die Beschäftigung mit den mündlichen Realisationsformen einer Sprache erst intensivierte, als die technischen Möglichkeiten zu ihrer Reproduzierbarkeit sich als handhabbar erwiesen. Kein Wunder also, dass die Gesprochene-Sprache-Forschung immer auch vom technischen Fortschritt beeinflusst wurde.

Die empirische Linguistik und der technische Fortschritt

Frühe Vorläufer der Gesprochenen-Sprache-Forschung waren die Dialektologie und die Phonetik, beide im 19. Jahrhundert entwickelt und beide sehr

früh auf die Nutzung technischer Hilfsmittel angewiesen. Eine Fülle von Beispielen für das Deutsche ließe sich anführen: in der Phonetik z.B. Pioniere wie Eduard Sievers seit den 1870er Jahren. Ein frühes Beispiel für Forschung an der Schnittstelle von Phonetik und Dialektologie ist die herausragende, kaum mehr bekannte schmale Habilitationsschrift von Theodor Frings über die »rheinische Accentuierung«.[2] Nach der Entwicklung der ersten Tonaufzeichnungsgeräte durch Thomas A. Edison (1877) und – heute sehr viel weniger bekannt, aber genauso einflussreich – Emil Berliner (1887) wurde sehr bald die einzigartige Chance erkannt, gesprochene Sprache zu konservieren. Um die Jahrhundertwende entstanden die ersten Tonarchive[3], im Vordergrund stand neben der Archivierung von Musik die Aufzeichnung von Reden berühmter Persönlichkeiten und auch schon die Dokumentation von Sprachen und Dialekten. Aber erst die durch Erfindung des Kondensatormikrofons 1927, die es ermöglichte, Töne aufzuzeichnen, ohne dass man in Trichter schreien musste, und durch die Entwicklung der Magnetbandtechnik zur komfortableren Aufzeichnung der Töne und durch die Produktion handhabbarer Heimtonbandgeräte nach dem 2. Weltkrieg wurde die Miniaturisierung der neuen Technologie eingeleitet, die ihren massenhaften und erschwinglichen Einsatz ermöglichte.[4]

In der deutschen Sprachwissenschaft war es insbesondere Eberhard Zwirner, der schon in den 1930er Jahren mit der Gründung des Deutschen Spracharchivs (DSAv) Weitsicht bewies und auch in den folgenden Jahrzehnten die Entwicklung unterstützender Technologien wahrnahm und für sprachwissenschaftliche Zwecke nutzte. Dazu gehört auch, dass in den 1970er Jahren erste Versuche stattfanden, die elektronische Datenverarbeitung für de Analysearbeiten des DSAv einzusetzen. Der Sprung vom analogen ins digitale Zeitalter wurde aber – nicht nur im DSAv – erst in den 1990er Jahren vollzogen und dieser Wechsel hat eine besondere Dimension, denn er markiert auf Grund der damit verbundenen völlig neuen Möglichkeiten einen Quantensprung in der Geschichte der empirischen Linguistik. Durch die Verschmelzung von Ton- und Computertechnik können Tonaufnahmen sehr viel schneller zugänglich gemacht und bearbeitet werden, sie können als Computerdateien gespeichert und als solche mit den verschiedensten Instrumenten der Computertechnik erschlossen werden. Auf dieser Basis ist auch für die Gesprochene-Sprache-Forschung die Nutzung sehr großer Datenmengen möglich, die für die Untersuchung von Besonderheiten und Gesetzmäßigkeiten der gesprochenen Sprache von großer Bedeutung sind.

Doch trotz aller technischen Neuerungen gilt auch hier, dass im Vergleich mit der Arbeit mit geschriebener Sprache die Beschaffung gesprochener Materialien sehr viel aufwändiger und letztlich auch kostenträchtiger ist. Das betrifft sowohl die Neuerhebung von Tondaten, auch wenn sie gleich mit digitaler Tontechnik erfolgt, und das betrifft insbesondere den Transfer vorhandener Daten, denn die Digitalisierung hochwertiger analoger Materialien muss von Experten vollzogen oder zumindest begleitet werden, um den Qualitätsstandard zu erhalten. Auf diese dringend erforderliche große Sorgfalt bei der Wahl der Methode und der Techniken muss gerade in Zeiten technologischen Wandels mit der damit verbundenen Aufbruchstimmung hingewiesen werden, um zu vermeiden, dass die neu gewonnenen Materialien vorzeitig und unerwartet auf digitalen Datenfriedhöfen landen.

Die DGD als Instrument zur Modernisierung des DSAv

In den 1990er Jahren stand dem Deutschen Spracharchiv eine Modernisierung ins Haus, weil die seit den 1950er Jahren verwendeten analogen Tonaufnahmen trotz professioneller Behandlung und Archivierung unaufhaltsam auf das Ende ihrer Lebensdauer zusteuerten.

Man entschied sich für Philosophie des »ewigen Datensatzes«, der – wenn erst einmal digital vorliegend – durch automatische Prozeduren von Generation zu Generation der digitalen Datenträger und Formate schnell und personalsparend portiert werden könnte. Nach gut zehn Jahren Erfahrung mit der Digitalisierung lässt sich die Richtigkeit (und Notwendigkeit) der damaligen Entscheidung nachdrücklich bestätigen. Für den zweiten Schritt der Modernisierung des DSAv, die Erstellung der Datenbank Gesprochenes Deutsch (DGD), die 1997 begonnen wurde, war die Digitalisierung der Archivalien notwendige Voraussetzung. Und mit der DGD wurden den Anforderungen und den Möglichkeiten, die sich mit dem Übergang in die digitale Welt ergaben, weitgehend entsprochen.[5]

Die DGD wurde im Rahmen eines von der Volkswagen-Stiftung geförderten Projekts entwickelt. Die Aufgabenstellung des Projekts zielte auf die »Computergestützte Erfassung und Erschließung der

Tonaufnahmen des Deutschen Spracharchivs zum gesprochenen Deutsch« (Projekttitel). Es diente dazu, das DSAv zu modernisieren und in eine virtuelle, über das Internet zugängliche Version zu überführen. Die DGD bietet im Internet abgestufte Zugänge zum virtuellen Archiv: Seit Februar 2002 ist die »Öffentliche Version« online, die (fast) alle Funktionen der DGD bietet und die volle Recherche in den Metadaten der Bestände erlaubt, aber nur wenig Archivmaterial wirklich nutzbar macht. Zurzeit bietet die Öffentliche Version den Zugang zu zehn zu Demonstrationszwecken ausgewählten Transkripten und Tonaufnahmen. Seit März 2003 ist die »Wissenschaftlerversion« im Internet erreichbar die den Zugang zu sämtlichen digital vorliegenden Materialien des DSAv bietet.

Die Bestände der DGD

Die Materialien der verschiedenen Korpora des DSAv lassen sich im Wesentlichen nach drei Materialtypen unterscheiden:

- Tonaufnahmen, die auf dem Massenspeichersystem der DGD als WAVE-Dateien liegen (und nur wegen der vielfach noch beschränkten Übertragungskapazitäten aus der Datenbank als datenreduzierte mp3- oder WMA-files heruntergeladen werden können);
- Transkripte, die überwiegend als Fließtexte vorliegen (auch bei den Dialektaufnahmen in einer standardsprachlichen Übertragung), aber auch in einer Partiturversion präsentiert werden, was insbesondere für die gesprächsanalytisch orientierten Korpora von Bedeutung ist;
- Metadaten, die allgemeine Informationen über die Korpora und im Einzelnen je nach Korpus unterschiedlich ausführliche Informationen zu den Sprecherinnen und Sprechern der Interaktionen, zur Situation der Aufnahme, zu den Inhalten und zum Stand der Aufbereitung der jeweiligen Materialien liefern.

Es ist wichtig, darauf hinzuweisen, dass die Materialien wegen der aufwändigen Aufbereitungsprozeduren in unterschiedlichem Maße für die DGD aufbereitet sind bzw. im Archiv vor—liegen. Zusammengefasst ergibt sich für die drei Materialstränge der folgende Stand der Aufbereitung für die DGD (Stand: Juni 2005)[6]: Die Datenbank enthält zurzeit

- dokumentarische Daten zu ca. 9400 Aufnahmen
- Transkripte zu mehr als 3100 Aufnahmen
- alignierte Transkripte aus fünf Korpora zu mehr als 1581 Aufnahmen, die also (ausschnittweise) auch online angehört werden können.

Die Funktionen der DGD

Die heterogenen Materialien des DSAv aus verschiedenen Erhebungs- und Forschungsprojekten erforderten vielseitig verwendbare und zukunftssichere digitale Archivformate und eine einheitliche Systematik. Mit den wichtigsten für die DGD entwickelten Werkzeugen zur Erschließung und Präsentation der Archivalien kann der Nutzer z.B.:

- mit Hilfe der Volltextrecherche in den Metadaten der Interaktionen suchen, d.h. er kann z.B. herausfinden, ob eine Interaktion mit einer Rechtsanwältin in einem Beratungsgespräch im Archiv vorliegt; oder eine Aufnahme im Dialekt von Altdorf oder Neustadt, oder eine Diskussion über den Sinn der Ehe; oder ein Gespräch mit einem Schmied aus Westfalen oder einer unter 50jährigen Lehrerin aus dem Schwäbischen; oder über Osterbräuche in Schlesien...;
- mit Hilfe der Volltextrecherche in den Transkripten suchen: nach einzelnen Wörtern, Phrasen oder Sätzen, nach Wortkombinationen und Kookurrenzen, die auch in Abständen auftreten können, z.B. nach »sowohl ... als auch« oder »ja ... aber« im Abstand von – sagen wir – zehn Wörtern...;
- durch vielfältige Optionen und Voreinstellungen in der Suchmaske und durch ihre Kombination Rechercheergebnisse gezielt einschränken oder ausweiten.
- vollständige Transkripte am Bildschirm lesen und bis zu 30sekündige Ausschnitte aus den dazu gehörigen Tonaufnahmen gleichzeitig anhören;
- die Ergebnisse der Suche in den Transkripten, die in einer KWIC-Liste (= **k**ey **w**ord **i**n **c**ontext) ausgegeben werden, anhören.
- die Metadaten und die Transkripte zu den Interaktionen herunterladen;
- die Tonaufnahmeausschnitte, die er durch Anklicken in den Transkripten erzeugt hat, herunterladen.

Diesen Nutzungsmöglichkeiten liegt die spezielle Aufbereitung der Töne und Texte des DSAv zu Grunde, die aligniert wurden, d.h. mit Hilfe eines im IDS weiter entwickelten und für die DGD modifizierten Verfahrens weitgehend automatisch synchronisiert wurden.

Archivzugang, Dienstleistungen und Perspektiven

Das DSAv war seit dem Abschluss der ersten größeren Aufnahmeaktionen in den 1960er Jahren eine in hohem Maße serviceorientierte Institution. 60 000 Kopien der analogen Tonaufnahmen und zahllose Kopien der Transkripte und der dokumentarischen Protokollbögen wurden in alle Welt verschickt.[7] Der Einsatz der DGD hat den Service auf eine völlig neue Basis gestellt: Kopien von Transkripten und Dokumentationen können vom registrierten Nutzer der Wissenschaftlerversion der DGD direkt heruntergeladen werden. Die Herstellung von Kopien der Tonaufnahmen kann für viele Korpora schon jetzt sehr schnell durchgeführt werden und wird in Zukunft vermutlich ebenfalls in der DGD direkt durch Herunterladen möglich sein.

Oben wurde erwähnt, dass die DGD über abgestufte Zugänge mit unterschiedlichen Rechten nutzbar ist. Völlig frei und anonym ist die Nutzung der Öffentlichen Version, die ein Nebenprodukt der Wissenschaftlerversion ist und in Zukunft für die Bedürfnisse der zufälligen und nicht vorinformierten Internetnutzer noch speziell angereichert und überarbeitet werden soll. Sie ist über die Internetadresse: http://dsav-oeff.ids-mannheim.de/DSAv zu erreichen. Auch die Anmeldung für die Wissenschaftlerversion erfolgt von dort über die Vergabe eines Zugangsnamens und Rücksendung eines Passworts per Email. In der Wissenschaftlerversion sind Recherchen in allen Materialien möglich, die digital vorliegen, einschließlich der Lesemöglichkeit von Transkripten am Bildschirm und des Abspielens von Tonaufnahmeausschnitten. Um diese Tonaufnahmen und die Transkripte auch herunterladen zu können, ist ein schriftlicher Antrag (per Post) auf Formularen erforderlich, die in der Wissenschaftlerversion bereit gestellt werden und verbunden mit einer Einverständniserklärung zu den Nutzungsbedingungen. Nötig sind diese abgestuften Zugangsmöglichkeiten auf Grund personen- und datenschutzrechtlicher Bestimmungen. Die Anmeldeprozeduren sind so einfach wie möglich gehalten. Zurzeit gibt es im Internet ca. 50000 Zugriffe jährlich auf die DGD, ca. 400 Nutzer haben sich für die Wissenschaftlerversion angemeldet und rund 50 Nutzer für die privilegierte Nutzung.

Anmerkungen

[1] Die Dialektologie hat sich in der ersten Hälfte des 20. Jahrhunderts systematisch dieser Methode bedient und dafür sogar einfache, schnell und kurrent zu schreibende Transkriptionssysteme entwickelt – für das Deutsche etwa die »Teuthonista«, vorgestellt 1924 in der ersten Ausgabe der gleichnamigen Zeitschrift, einer Vorläuferin der heutigen »Zeitschrift für Dialektologie und Linguistik«. Vgl. auch Wagener (1988); S. 118/119.

[2] Frings (1916).

[3] Für einen groben Überblick über die Geschichte der Tonarchive aus der Sicht eines Sprachwissenschaftlers s. Wagener (1988), S. 164f.

[4] Auch diese technologische Entwicklung ist übrigens erst durch ihre militärische Nutzung im 2. Weltkrieg entscheidend forciert worden.

[5] Ausführlicher dazu vgl. Wagener (2002).

[6] Detaillierte Informationen zum aktuellen Stand der Aufbereitung der Materialien lassen sich der »Bestandsübersicht« entnehmen, die in der DGD von den Korpusseiten aus erreichbar ist (oder direkt unter der Adresse: http://dsav-wiss.ids-mannheim.de/DSAv/KORPORAB.HTM).

[7] Genauere Informationen dazu finden sich bei Bethge (1976) und Knetschke/Sperlbaum (1983).

Literatur

Bethge, Wolfgang (1976): Vom Werden und Wirken des Deutschen Spracharchivs. In: Zeitschrift für Dialektologie und Linguistik 43, S. 22-53.

Frings, Theodor (1916): Die rheinische Accentuierung. Marburg (= Deutsche Dialektgeographie 14).

Knetschke, Edeltraud/Sperlbaum, Margret (1983): Das Deutsche Spracharchiv im Institut für deutsche Sprache. 2. Aufl., Mannheim (= IDS-Mitteilungen 6).

Wagener, Peter (1988): Untersuchungen zur Methodologie und Methodik der Dialektologie. Marburg (= Deutsche Dialektgeographie 86).

Wagener, Peter (2002): Gesprochenes Deutsch online. Zur Modernisierung des Deutschen Spracharchivs. In: Zeitschrift für Dialektologie und Linguistik 69/3, S. 314-335.